Assignment #1: Data Integration

1. Objectives

The objective of this project is to apply some of the topics learned this semester to real-world problems. The data management component aims to create an analytical dataset containing sufficient information for the exploratory data analysis (EDA) to be discussed next week. The specific variables and details to be included in the final dataset will be outlined in subsequent sections.

2. Data Sources

The datasets involved in this project are as follows:

- Presidential Election Data: [`countypresidential_election_2000-2020.csv`]
- Unemployment Data: [`Unemployment.csv`]
- Poverty Data: [`PovertyEstimates.csv`]
- Education Data: [`Education.csv`]

3. Specific Requirements for Data Preparation

- **Presidential Election Data:** The final analytical dataset should include the following information:
  - Presidential Election Data
  - Only use 2020 election data.
  - Retain data for the two major parties (Democrats and Republicans).
  - Aggregate the total votes and keep the winning party in the dataset.
  - Specific variables include:
    - County FIPS code
    - State name
    - County name
    - Total votes received by the winning party
    - Name of the winning party

- **Unemployment Data:** Keep only 2020 unemployment data (or the most recent year if 2020 data is unavailable, retaining only one record per county. Specific variables include:
  - County FIPS code
  - Unemployment rate

- **Poverty Data**: Keep only the 2019 poverty rate (variable name: `PCTPOVALL_2019`). Specific variables include:
  - County FIPS code

- **Education Data**: Keep only the percentage of education levels (2015–2019). Education levels to include:
  - Less than a high school diploma
  - High school diploma only
  - Some college (1–3 years)
  - Four-year college degree or higher

  Also include:
  - County FIPS code

After processing these datasets, merge them into a single dataset where each county has only one record.

4. Suggestions on Coding Conventions

Below are some best practices for coding projects with multiple tasks:

- **Formatting**
  - Avoid placing multiple statements on one line (reduces readability).
  - List each variable on a separate line when declaring multiple variables.
  - Use indentation (e.g., in SAS, indent all statements within a `DATA` or `PROC` step except the first and last lines).
- **Commenting**
  - Include a program header with key details (author, purpose, date, etc.).
  - Add comments when modifying code to explain changes.
  - Comment on non-obvious logic or clever solutions for clarity.

- **Naming Conventions**
  - Dataset names should be descriptive (e.g., `election_2020_clean`).
  - Variable names should be clear and meaningful.

Summary of the Integrated Data Set

The resulting dataset contains [X] rows and [Y] columns, providing a structured foundation for analysis. Below is a summary of its key details:

- Size: [X] rows × [Y] columns
- Variables and Data Types:
  - [Variable 1]: [Data type – e.g., numeric, categorical, datetime]
  - [Variable 2]: [Data type]
  - [Variable 3]: [Data type]

  (Continue listing all variables as needed.)

Note: This dataset will be used in next week's exploratory data analysis (EDA) assignment, which will involve summarizing distributions, detecting missing values, and examining relationships between variables.

Submission Components

1. The summary of the integrated data set.
2. Code (SAS or R) used in data integration.
3. Export the data set as csv file and upload it to your GitHub repository.
4. Provide the URL of this data set.